United States
Environmental Protection
Agency

Office of Water
Office of Science and Technology
Washington, DC 20460

EPA-822-R-07-004
September 2007
www.epa.gov

Nutrient Criteria
Technical Guidance Manual

Wetlands

Full Guidance Document available in full at:
http://www.epa.gov/waterscience/criteria/nutrient/guidance/

Chapter 6 Database Development and New Data Collection

6.1 INTRODUCTION

A database of relevant water quality information can be an invaluable tool to States as they develop nutrient criteria. In some cases, existing data are available and can provide additional information that is specific to the region where criteria are to be set. However, little or no data are available for most regions or parameters and creating a database of newly gathered data is strongly recommended. In the case of existing data, the data should be located and their suitability (type and quality and sufficient associated metadata) ascertained. It is also important to determine how the data were collected to ensure that future monitoring efforts are compatible with earlier approaches.

Databases operate much like spreadsheet applications but have greater capabilities. Databases store and manage large quantities of data and allow viewing and exporting of data sorted in a variety of ways, while spreadsheets analyze and graphically display small quantities of data. Databases can be used to organize existing information, store newly gathered monitoring data, and manipulate data for water quality criteria development. Databases can sort data for export into statistical analyses programs, spreadsheets, and graphics programs. This chapter will discuss the role of databases in nutrient criteria development and provide a brief review of existing sources of nutrient-related water quality information for wetlands.

6.2 DATABASES AND DATABASE MANAGEMENT

A database is a collection of information related to a particular subject or purpose. Databases are arranged so that individual values are kept separate, yet can be linked to other values based on some common denominator (such as association of time or location). Geographic Information Systems (GIS) are geo-referenced relational databases that have a geographical component (i.e., spatial platform) in the user interface. Spatial platforms associated with a database allow geographical display of sets of sorted data. GIS platforms such as ArcView™, ArcInfo™, and MapInfo™ are frequently used to integrate spatial data with monitoring data for watershed analysis. Data stored in simple tables, relational databases, or geo-reference databases can also be located, retrieved, and manipulated using queries. A query allows the user to find and retrieve only the data that meets user-specified conditions. Queries can also be used to update or delete multiple records simultaneously and to perform built-in or custom calculations of data. Data in tables can be analyzed and printed in specific layouts using reports. Data can be analyzed or presented in a specific way in print by creating a report. The most effective use of these tools requires a certain amount of training, expertise, and software support, especially when using geo-referenced data.

To facilitate data storage, manipulation, and calculations, it is highly recommended that historical and present-day data be transferred to a relational database (i.e., Access™). Relational databases store data in tables as sets of rows and columns and are powerful tools for data manipulation and initial data reduction. They allow selection of data by specific, multiple criteria and definition and redefinition of linkages among data components. Data queries can also be exported to GIS, provided that the data is related to some geo-referenced coordinate system.

POTENTIAL DATA SOURCES

EPA Water Quality Data

STORET
EPA has many programs of national scope that focus on collection and analysis of water quality data. The following presents information on several of the databases and national programs that may be useful to water quality managers as they compile data for criteria development. STORET STOrage and RETrieval system (STORET) is EPA's national database for water quality and biological data.

Environmental Monitoring and Assessment Program (EMAP)
The Environmental Monitoring and Assessment Program is an EPA research program designed to develop the tools necessary to monitor and assess the status and trends of national ecological resources (see EMAP Research Strategy on the EMAP Web site: www.epa.gov/emap). EMAP's goal is to develop the scientific understanding for translating environmental monitoring data from multiple spatial and temporal scales into assessments of ecological condition and forecasts of future risks to the sustainability of the Nation's natural resources. Data from the EMAP program can be downloaded directly from the EMAP Web site (www.epa.gov/emap/ . The EMAP Data Directory contains information on available data sets, including data and metadata (language that describes the nature and content of data). Current status of the data directory, as well as composite data and metadata files, are available on this Web site.

U.S. Geological Surve) (USGS) Water Data

The USGS has national and distributed databases on water quantity and quality for waterbodies across the nation. Much of the data for rivers and streams are available through the National Water Information System (NWIS). These data are organized by State, Hydrologic Unit Codes (HUCs), latitude and longitude, and other descriptive attributes. Most water quality chemical analyses are associated with an instantaneous streamflow at the time of sampling and can be linked to continuous streamflow to compute constituent loads or yields. The most convenient

method of accessing the local databases is through the USGS State representative. Every State office can be reached through the USGS home page at: http://www.usgs.gov.

HBN and NASQAN

USGS data from several national water quality programs covering large regions offer highly controlled and consistently collected data that may be particularly useful for nutrient criteria analysis. Two programs, the Hydrologic Benchmark Network (HBN) and the National Stream Quality Accounting Network (NASQAN), include routine monitoring of rivers and streams over the past 30 years. The HBN consists of 63 relatively small, minimally disturbed watersheds. HBN data were collected to investigate naturally-induced changes in streamflow and water quality and the effects of airborne substances on water quality. The NASQAN program consists of 618 larger, more culturally influenced watersheds. NASQAN data provides information for tracking water-quality conditions in major U.S. rivers and streams. The watersheds in both networks include a diverse set of climatic, physiographic, and cultural characteristics. Data from the networks have been used to describe geographic variations in water-quality concentrations, quantify water-quality trends, estimate rates of chemical flux from watersheds, and investigate relations of water quality to the natural environment and anthropogenic contaminant sources.

WEBB

The Water, Energy, and Biogeochemical Budgets (WEBB) program was developed by USGS to study water, energy, and biogeochemical processes in a variety of climatic/regional scenarios. Five ecologically diverse watersheds, each with an established data history, were chosen. This program may prove to be a rich data source for ecoregions in which the five watersheds are located. Many publications on the WEBB project are available. See the USGS Web site for more details (http://water.usgs.gov/nrp/webb/about.html).

US Department of Agriculture (USDA)
Agricultural Research Service (ARS)

The USDA ARS houses the Natural Resources and Sustainable Agricultural Systems Scientific Directory (http://hydrolab.arsusda.gov/arssci.html), which has seven national programs to examine the effect of agriculture on the environment. The program on Water Quality and Management addresses the role of agriculture in nonpoint source pollution through research on Agricultural Watershed Management and Landscape Features, Irrigation and Drainage Management Systems, and Water Quality Protection and Management Systems. Research is conducted across the country and several models and databases have been developed. Information on research and program contacts is listed on the Web site (http://www.nps.ars.usda.gov/programs/nrsas.htm).

Forest Service

The Forest Service has designated research sites across the country, many of which are Long Term Ecological Research (LTER) sites. Many of the data from these experiments are available in the USFS databases located on the Web site (http://www.fs.fed.us/research/). Most of the data are forest-related but may be of use for determining land uses and questions on silviculture runoff.

National Science Foundation (NSF)

The National Science Foundation (NSF) funds projects for the Long Term Ecological Research (LTER) Network. The Network is a collaboration of over 1,100 researchers investigating a wide range of ecological topics at 24 different sites nationwide. The LTER research programs are not only an extremely rich data source, but also a source of data available to anyone through the Network Information System (NIS), the NSF data source for LTER sites. Data sets from sites are highly comparable due to standardization of methods and equipment.

U.S. Army Corps of Engineers (COE)

The U.S. Army Corps of Engineers (COE) is responsible for many federal wetland jurisdiction issues. Although a specific network of water quality monitoring data does not exist, specific studies on wetlands by the COE may provide suitable data. The COE focuses more on water quantity issues than on water quality issues. As a result, much of the wetland system data collected by the COE does not include nutrient data. Nonetheless, the COE does have a large water sampling network and supports USGS and EPA monitoring efforts in many programs. A list of the water quality programs that the COE actively participates in can be found at http://www.usace.army.mil/public.html.

U.S. Department of the Interior, Bureau of Reclamation (BuRec)

The Bureau of Reclamation of the U.S. Department of the Interior manages many irrigation and water supply reservoirs in the West, some of which may have wetland applicable data available. These data focus on water supply information and limited water quality data. However, real time flow data are collected for rivers supplying water to BuRec, which may be useful if a flow component of criteria development is chosen. These data can be gathered on a site-specific basis from the BuRec Web site: http://www.usbr.gov.

State Monitoring Programs

Some States may have wetland water quality data as part of a research study, use attainability analysis (UAA), or to assess mitigation or nutrient related impacts. Most of this data is collected by State natural resources or environmental protection agencies, or by regional water management authorities. Data collected by State water quality monitoring programs can be used for nutrient criteria development and may provide pertinent data sources, although they may be regionally limited. These data should be available from the agencies responsible for monitoring.

Volunteer Monitoring Programs

State and local agencies may use volunteer data to screen for water quality problems, establish trends in waters that would otherwise be unmonitored, and make planning decisions. Volunteers benefit from learning more about their local water resources and identifying what conditions or activities might contribute to pollution problems. As a result, volunteers frequently work with clubs, environmental groups, and State or local governments to address problem areas. The EPA supports volunteer monitoring and local involvement in protecting our water resources.

Academic and Literature Sources

Most of the data available on water and soil quality in wetlands is the result of research studies conducted by academic institutions. Much of the research conducted by the academic community, however, was not conducted for the purpose of spatial or long-term biogeochemical characterization of the nation's wetlands; instead, water quality information was often collected to characterize the environmental conditions under which a particular study or experiment was conducted. Infrequently, spatial studies of limited extent or duration were conducted. Data collected from these sources, therefore, may not be sufficiently representative of the population of wetlands within an ecoregion. However, this limited data may be the only information available and therefore could be useful for identifying reference conditions or determining where to begin a more comprehensive survey to support development of nutrient criteria. Academic research data is available from researchers and the scientific literature.

6.3 QUALITY OF HISTORICAL AND COLLECTED DATA

The value of older historical data is a recurrent problem because data quality is often unknown. Knowledge of data quality is also problematic for long-term data repositories such as STORET and long-term State databases, where objectives, methods, and investigators may have changed many times over the years. The most reliable data tend to be those collected by a single agency using the same protocol. Supporting documentation should be examined to determine the consistency of sampling and analytical protocols. The suitability of data in large, heterogeneous data repositories for establishing nutrient criteria are described below. These same factors need to be taken into account when developing a new database such that future investigators will have sufficient information necessary to evaluate the quality of the database.

LOCATION

Geo-referenced data is extremely valuable in that it allows for aggregating and summarizing data according to any GIS coverage desired, whether the data was historically related to a particular coverage theme or not. However, many studies conducted prior to the availability and accuracy of hand held Global Positioning System (GPS) units relied on narrative and less definitive descriptions of location such as proximity to transportation corridor, county, or nearest municipal center. This can make comparison of data, depending upon desired spatial resolution, difficult. Knowledge of the rationale and methods of site selection from the original investigators may supply valuable information for determining whether inclusion of the site or study in the database is appropriate based on potential bias relative to overall wetland data sources. STORET and USGS data associated with the National Hydrography Dataset (NHD) are geo-referenced with latitude, longitude, and Reach File 3 (RF3) codes (http://nhd.usgs.gov/). In addition, STORET often contains a site description to supplement location information. Metadata of this type, when known, is frequently stored within large long-term databases.

VARIABLES AND ANALYTICAL METHODS

Each separate analytical method yields a unique variable. For example, five ways of measuring TP result in five unique variables. Data generated using different analytical methods should not be combined in data analyses because methods differ in accuracy, precision, and detection limits. Data generated from one method may be too limited, making it important to select the most frequently used analytical methods in the database. Data that were generated using the same analytical methods may not always be obvious because of synonymous names or analytical methods. Consistency in taxonomic conventions and indicator measurements is likewise important for biological variables and multimetric indices comparisons. Review of recorded data and analytical methods by knowledgeable personnel is important to ensure that there are no problems with data sets developed from a particular database.

LABORATORY QUALITY CONTROL (QC)

Data generated by agencies or laboratories with known quality control/quality assurance protocols are most reliable. Laboratory QC data (blanks, spikes, replicates, known standards) are infrequently reported in larger data repositories. Records of general laboratory quality control protocols and specific quality control procedures associated with specific data sets are valuable in evaluating data quality. However, premature elimination of lower quality data can be counterproductive because the increase in variance caused by analytical laboratory error may be negligible compared to natural variability or sampling error, especially for nutrients and related water quality parameters. However, data of uncertain and undocumented quality should not be accepted.

Water column nutrient data can be reported in different units, e.g., ppm, mg/L, mmoles. Reporting of nutrient data from other strata such as soils, litter, and vegetation can further expand the list of reporting units (e.g., mg/kg, g/kg, %, mg/cm^3). In many instances, conversion of units is possible; however, in other instances unit conversion is not possible or is lacking support information for conversion. Consistency in reporting units and the need to provide conversion tables cannot be overemphasized.

DATA COLLECTING AGENCIES

Selecting data from particular agencies with known, consistent sampling and analytical methods and known quality will reduce variability due to unknown quality problems. Requesting data review for quality assurance from the collecting agency will reduce uncertainty about data quality.

TIME PERIOD

Long-term records are critically important for establishing trends. Determining if trends exist in the time series database is also important for characterizing reference conditions for nutrient criteria. Length of time series data needed for analyzing nutrient data trends is discussed in Chapter 7.

INDEX PERIOD

An index period—the time period most appropriate for sampling—for estimating average concentrations can be established if nutrient and water quality variables were measured through seasonal cycles. The index period may be the entire year or the summer growing season. The best index period is determined by considering wetland characteristics for the region, the quality and quantity of data available, and estimates of temporal variability (if available). Consideration of the data available relative to longer-term oscillations in environmental conditions (e.g., dry years, wet years) should also be taken into account such that the data is representative and appropriate. Additional information and considerations for establishing an index period are discussed in Chapter 7.

REPRESENTATIVENESS

Data may have been collected for specific purposes. Data collected for toxicity analyses, effluent limit determinations, or other pollution problems may not be useful for developing nutrient criteria. Further, data collected for specific purposes may not be representative of the region or wetland classes of interest. The investigator should determine if all wetlands or a subset of the wetlands in the database are representative of the population of wetlands to be characterized. If a sufficient sample of representative wetlands cannot be found, then a new survey is strongly recommended.

6.4 COLLECTING NEW DATA

New data should be collected when no data presently exist or the data available are not suitable, and should be gathered following the sampling design protocols discussed in Chapter 4. New data collection activities for developing nutrient criteria should focus on filling in gaps in the database and collecting spatially representative regional monitoring data. In many cases, this may mean starting from scratch because no data presently exists or the data available are not suitable. Data gathered under new monitoring programs should be imported into databases or spreadsheets and, if comparable, merged with existing data for criteria development. It is best to archive the data with as much data-unique information (meta-data) as possible. It is always possible to aggregate at a later time, but impossible to separate lumped data without having the parameter needed to partition the data set. Redundancy may also be a problem but can more easily be avoided when common variables or parameters are kept in each database (i.e., dates may be very important). The limitations and qualifications of each data set should be known and data 'tagged', if possible, before combining them. The following five factors should be considered when collecting new data and before combining new data with existing data sets: representativeness, completeness, comparability, accuracy, and precision.

REPRESENTATIVENESS

Sampling program design (when, where, and how you sample) should produce samples that are *representative* or typical of the regional area being described and the classes of wetlands present. Sampling designs for developing nutrient criteria are addressed in Chapter 4. Databases populated by data from the literature or historical studies will not likely provide sufficient spatial or class representation of a region. Data interpretation should recognize these gaps and be limited until gaps are filled using additional survey information.

COMPLETENESS

A QA/QC plan should describe how to complete the data set in order to answer questions posed (with a statistical test of given power and confidence) and the precautions being taken to ensure that completeness. Data collection procedures should document the extent to which these conditions have been met. Incomplete data sets may not invalidate the collected data but may reduce the rigor of statistical analyses. Precautions to ensure completeness may include collecting extra samples, having back-up equipment in the field, copying field notebooks after each trip, and/or maintaining duplicate sets of data in two locations.

COMPARABILITY

In order to compare data collected under different sampling programs or by different agencies, sampling protocols and analytical methods should demonstrate comparable data. The most efficient way to produce comparable data is to use sampling designs and analytical methods that are widely used and accepted and examined for compatibility with other monitoring programs prior to initiation of a survey. Comparability should be assessed for field sample collection, sample preservation, sample preparation and analysis, and among laboratories used for sample analyses.

ACCURACY

To assess the accuracy of field instruments and analytical equipment, a standard (a sample with a known value) should be analyzed and the measurement error or bias determined. Internal standards should periodically be checked with external standards provided by acknowledged sources. At Federal, State, and local government levels, the National Institute of Standards and Technology (NIST) provides advisory and research services to all agencies by developing, producing, and distributing standard reference materials for vegetation, soils, and sediments. Standards and methods of calibration are typically included with turbidity meters, pH meters DO meters, and DO testing kits. The U.S. EPA, USGS, and some private companies provide reference standards or QC samples for nutrients.

VARIABILITY

The variability in field measurements and analytical methods should be demonstrated and documented to identify the source and magnitude of variability when possible. EPA QA/QC guidance provides an explanation and protocols for measuring sampling variability (USEPA 1998c).

DATA REDUCTION

For data reduction, it is important to have a clear idea of the analysis that will be performed and a clear definition of the sample unit for analysis. For example, a sample unit might be defined as “a wetland during July- August.” For each variable measured, a mean value would then be estimated for each wetland during the July-August index period on record. Analyses are then conducted on the observations (estimated means) for each sample unit, not with the raw data. Steps recommended for reducing the data include:

1. Selecting the long-term time period for analysis;

2. Selecting an index period;

3. Selecting relevant variables of interest;

4. Identifying the quality of analytical methods;

5. Identifying the quality of the data recorded; and,

6. Estimating values for analysis (mean, median, minimum, maximum) based on the reduction selected.

6.5 QUALITY ASSURANCE / QUALITY CONTROL (QA/QC)

The validity and usefulness of data depend on the care with which they were collected, analyzed, and documented. EPA provides guidance on data quality assurance (QA) and quality control (QC) (USEPA 1998c) to assure the quality of data. Factors that should be addressed in a QA/QC plan are elaborated below. The QA/QC plan should state specific goals for each factor and should describe the methods and protocols used to achieve the goals.

1. Who will use the data?
2. What the project's goals/objectives/questions or issues are?
3. What decision(s) will be made from the information obtained?
4. How, when, and where project information will be acquired or generated?
5. What possible problems may arise and what actions can be taken to mitigate their impact on the project?
6. What type, quantity, and quality of data are specified?
7. How "good" those data have to be to support the decision to be made?
8. How the data will be analyzed, assessed, and reported?